ANALYSIS OF RETAIL TRANSACTIONS USING GAUSSIAN MIXTURE MODELS IN A DATA MINING SYSTEM

CROSS REFERENCE TO RELATED APPLICATIONS

This application is related to the following co-pending and commonly assigned patent applications:

Application Serial No. xx/xxx,xxx, filed on same date herewith, by Paul M. Cereghini and Scott W. Cunningham, and entitled "ARCHITECTURE FOR A DISTRIBUTED RELATIONAL DATA MINING SYSTEM," attorneys' docket number 9141;

Application Serial No. xx/xxx,xxx, filed on same date herewith, by Scott W. Cunningham, and entitled "IMPROVEMENTS TO GAUSSIAN MIXTURE MODELS IN A DATA MINING SYSTEM," attorneys' docket number 9143; and

Application Serial No. xx/xxx,xxx, filed on same date herewith, by Mikael Bisgaard-Bohr and Scott W. Cunningham, and entitled "DATA MODEL FOR ANALYSIS OF RETAIL TRANSACTIONS USING GAUSSIAN MIXTURE MODELS IN A DATA MINING SYSTEM," attorneys' docket number 9684;

all of which applications are incorporated by reference herein.

BACKGROUND OF THE INVENTION

1. Field of the Invention.

This invention relates to a computer-implemented data mining system, and in particular, to a system for analyzing retail transactions using Gaussian Mixture Models in a distributed relational data mining system.

2. Description of Related Art.

Many computer-implemented systems are used to analyze commercial and financial transaction data. In many instances, such data is analyzed to gain a better understanding of customer behavior by analysis of customer transactions.

Prior art methods for analyzing customer transactions often involve one or more of the following techniques:

1. Ad hoc querying: This methodology involves the iterative analysis of transaction data by human effort, using querying languages such as SQL.

2. On-line Analytical Processing (OLAP): This methodology involves the application of automated software front-ends that automate the querying of relational databases storing transaction data and the production of reports therefrom.

3. Statistical packages: This methodology requires the sampling of transaction data, the extraction of the data into flat file or other proprietary formats, and the application of general purpose statistical or data mining software packages to the data.

Nonetheless, there remains a need for improved techniques for analyzing transaction data.

SUMMARY OF THE INVENTION

A computer-implemented data mining system that analyzes data using Gaussian Mixture Models. The data is accessed from a database, and then an Expectation-Maximization (EM) algorithm is performed in the computer-implemented data mining system to create the Gaussian Mixture Model for the accessed data. The EM algorithm generates an output that describes clustering in the data by computing a mixture of probability distributions fitted to the accessed data.

BRIEF DESCRIPTION OF THE DRAWINGS

Referring now to the drawings in which like reference numbers represent corresponding parts throughout:

FIG. 1 illustrates an exemplary hardware and software environment that could be used with the present invention;

FIG. 2 is a diagram that illustrates the structure of a data model according the preferred embodiment of the present invention; and

FIG. 3 is a flowchart that illustrates the logic for crating and using the data model 200 according the preferred embodiment of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

In the following description of the preferred embodiment, reference is made to the accompanying drawings which form a part hereof, and in which is shown by way of illustration a specific embodiment in which the invention may be practiced. It is to be understood that other embodiments may be utilized and structural changes may be made without departing from the scope of the present invention.

OVERVIEW

The present invention represents a way of producing customer segments from a transactional database. A segment is a grouping of data elements organized about one or more attributes. These customer segments may serve as the basis for merchandising or marketing campaigns. They are a powerful basis for analysis of customer behavior, and they are useful means for summarizing the often exhaustive contents of transaction-based data warehouses.

HARDWARE AND SOFTWARE ENVIRONMENT

FIG. 1 illustrates an exemplary hardware and software environment that could be used with the present invention. In the exemplary environment, a computer system 100 implements a data mining system in a three-tier client-server architecture comprised of a first client tier 102, a second server tier 104, and a third server tier 106. In the preferred embodiment, the third server tier 106 is coupled via a network 108 to one or more data servers 110A-110E storing a relational database on one or more data storage devices 112A-112E.

The client tier 102 comprises an Interface Tier for supporting interaction with users, wherein the Interface Tier includes an On-Line Analytic Processing (OLAP) Client 114 that provides a user interface for generating SQL statements that retrieve data from a database, an Analysis Client 116 that displays results from a data mining algorithm, and an Analysis Interface 118 for interfacing between the client tier 102 and server tier 104.

The server tier 104 comprises an Analysis Tier for performing one or more data mining algorithms, wherein the Analysis Tier includes an OLAP Server 120 that schedules and prioritizes the SQL statements received from the OLAP Client 114, an Analysis Server 122 that schedules and invokes the data mining algorithm to analyze the data retrieved from the database, and a Learning Engine 124 performs a Learning step of the data mining algorithm. In the preferred embodiment, the data mining algorithm comprises an Expectation-Maximization procedure that creates a Gaussian Mixture Model using the results returned from the queries.

The server tier 106 comprises a Database Tier for storing and managing the databases, wherein the Database Tier includes an Inference Engine 126 that performs an Inference step of the data mining algorithm, a relational database management system (RDBMS) 132 that performs the SQL statements against a Data Mining View 128 to retrieve the data from the database, and a Model Results Table 130 that stores the results of the data mining algorithm.

The RDBMS 132 interfaces to the data servers 110A-110E as mechanism for storing and accessing large relational databases. The preferred embodiment comprises the Teradata® RDBMS, sold by NCR Corporation, the assignee of the present invention, which excels at high volume forms of analysis. Moreover, the RDBMS 132 and the data servers 110A-110E may use any number of different parallelism mechanisms, such as hash partitioning, range partitioning, value partitioning, or other partitioning methods. In addition, the data servers 110 perform operations against the relational database in a parallel manner as well.

Generally, the data servers 110A-110E, OLAP Client 114, Analysis Client 116, Analysis Interface 118, OLAP Server 120, Analysis Server 122, Learning Engine 124, Inference Engine 126, Data Mining View 128, Model Results Table 130, and/or RDBMS 132 each comprise logic and/or data tangibly embodied in and/or accessible from a device, media, carrier, or signal, such as RAM, ROM, one or more of the data storage devices 112A-112E, and/or a remote system or device communicating with the computer system 100 via one or more data communications devices.

However, those skilled in the art will recognize that the exemplary environment illustrated in FIG. 1 is not intended to limit the present invention. Indeed, those skilled in the art will recognize that other alternative environments may be used without departing from the scope of the present invention. In addition, it should be understood that the present invention may also apply to components other than those disclosed herein.

For example, the 3-tier architecture of the preferred embodiment could be implemented on 1, 2, 3 or more independent machines. The present invention is not restricted to the hardware environment shown in FIG. 1.

OPERATION OF THE DATA MINING SYSTEM

The present invention allows analysts to gain a better understanding of customer behavior by means of a thorough cluster analysis of customer transactions, although customer identification is not required for the analysis. The goal of cluster analysis is to group items coherently according to perceived similarities in the data.

Gaussian Mixture Models are the particular form of clustering that is used in the analysis performed by the present invention. The data for the clustering consists of customer transactions or "baskets." The baskets are grouped according to behavioral similarities revealed during shopping. The resulting transaction clusters offer an insight into the shopping behavior of both individuals and groups. Marketing professionals call these clusters "customer segments."

When applied to basket data, clustering provides three broad opportunities for analysis and business improvement. Of primary importance in all these analyses is the economic impact of the customer segment.

1. Price and Promotion Analysis: How responsive are various segments to the pricing and promotion of products? What product attributes are most appealing to each segment? Which segments are brand loyal or prefer store brands?

2. Demographic and Locational Analysis: What are the demographic characteristics of customer segments? How do store formats and locations affect the mix of customer segments? How is the mix of customer segments changing over time?

3. Purpose and Interest Analysis: What was the apparent purpose of the visit? What departments were visited? How much variety in shopping was displayed by customer segment? How frequently did the customer shop? Which items satisfy particular shopping needs?

Demographic data is useful since it allows knowledge about particular customers to be extended to representative customer segments. This is used, for example, in the demographic typing of customer segments. It is also used in establishing shopping frequency statistics by customer segment.

Affinity is a form of analysis that examines the frequency with which various products are purchased both together and separately. Segmentation reveals the very different patterns of purchases and affinities that are possible across distinct customer groups. Segmentation therefore is a powerful extension to standard affinity analysis.

There are many ways of grouping transactions to analyze customer behavior. In addition, many forms of customer analysis deal with summary data about customers as a whole. The advantages of using Gaussian Mixture Models, a statistical form of analysis, are four-fold:

1. Automation: Gaussian Mixture Models are automated statistical procedures suitable for finding patterns and clusters in databases. As a result, machine techniques can be applied to the searching and scanning of databases, thereby relieving human analysts of the task.

2. Statistical Quality: Gaussian Mixture Models find robust and repeatable patterns in the database. In addition, there is an intrinsic measure of model quality, known as the "log-likelihood." This allows users to interpret the quality of the results and to explicitly examine shortcomings of the solutions.

3. Summarization: Gaussian Mixture Models provide effective summarization of exhaustive databases of customer transactions. The resulting summary allows analysts to deal with the most representative transactions in the database.

4. Disaggregation: By separating sources of variability in the transaction database, analysts gain a better understanding of how customer behavior varies. Armed

with this knowledge, retailers may act upon distinct customer groups to encourage profitable behavior.

There are two components to the present invention: a data model generated by the data mining system 100 and an algorithm performed by the data mining system 100 to create the data model. The data model comprises a Gaussian Mixture Model that stores transactional data and provides a minimum specification for the transactional data needed in the analysis. The algorithm performs the mapping function necessary to create the data model by aggregating the transactional data for cluster analysis. The result, as noted, is a grouping of the transactional data into segments, wherein each segment may be summarized by a set of prototypical behaviors.

The preferred embodiment of the present invention provides a number of advantages.

- First, the present invention is entirely automated, requiring few arbitrary decisions or expectations regarding the solution or structure on the part of the analyst, which differs substantially from "ad hoc querying" used in prior efforts.
- Second, the present invention employs "fuzzy sets" that result in high fidelity reproduction and summarization of database results, which differs substantially from prior efforts, such as OLAP systems that utilize SQL sets as a means of defining customer segments.
- Third, the present invention uses a single, dedicated algorithm with a well-defined data model. As a result, the present invention requires very little specialized knowledge to utilize and interpret the results. This represents a significant difference from prior designs utilizing statistical packages.

DATA MODEL

FIG. 2 is a diagram that illustrates the structure of a data model 200 according the preferred embodiment of the present invention. The data model 200 comprises a Gaussian Mixture Model, and may be stored in the relational database managed by the

RDBMS 132. The data model 200 is a structured way of storing transactional data. This transactional data might be obtained, for example, from a point-of-sale device.

In the preferred embodiment, three tables are used in the model 200: a basket table 202, an item table 204 and a department table 206. The basket data 202 contains summary information about transactions. The item table 204 contains information about individual items purchased by customers. The department table 206 is a source of useful aggregate information about transaction sales by store department (although this data may ultimately be derived entirely from the item table 204).

This data is then mapped into a single flat table format, perhaps using a database view, to produce the correct level of aggregation for the statistical analysis. The analysis requires one row to one customer transaction. Multiple transactions by the same customer are not of concern. In general, customers can not be uniquely identified from this format or view.

ALGORITHM

FIG. 3 is a flowchart that illustrates the logic for creating and using the data model 200 according the preferred embodiment of the present invention.

Block 300 represents the transactional data being accessed and retrieved from the relational database by the RDBMS 132.

Block 302 represents a Gaussian Mixture Model algorithm being applied to the transactional data by the Analysis Server 122,the Learning Engine 124, and the Inference Engine 126 to create the data model 200. The Gaussian Mixture Model assumes that the transactions result from a mix of distinct customer behaviors.

Gaussian Mixture Models are a form of machine learning, described in more detail in sources such as Roweis, S.T. and Ghahramani, Z. (1999), A Unifying Review of Linear Gaussian Models, Neural Computation 11(2):305-345, which publication is incorporated by reference herein. One implementation of an algorithm for generating the Gaussian Mixture Models is described in co-pending and commonly-assigned Application Serial No. xx/xxx,xxx, filed on same date herewith, by Scott W. Cunningham, and entitled "IMPROVEMENTS TO GAUSSIAN MIXTURE MODELS

IN A DATA MINING SYSTEM," attorneys' docket number 9143, which application is incorporated by reference herein.

Block 304 represents behavioral "profiles" reported across a range of selected variables being returned from the data model 200 maintained by the Analysis Server 122 to the Analysis Client 116.

Block 306 represents a range of behaviors expected from each variable, in each cluster, being returned from the data model 200 maintained by the Analysis Server 122 to the Analysis Client 116.

Block 308 represents the relative mix or proportions of behaviors in the database being returned from the data model 200 maintained by the Analysis Server 122 to the Analysis Client 116.

Block 310 represents an assignment of analyzed transactions to associated customer behaviors being returned from the data model 200 maintained by the Analysis Server 122 to the Analysis Client 116. The default results show the mixes of behaviors represented within any given transaction. Alternatively, the results can be formatted so that one transaction has one, and only one, associated behavior. (This "winner-takes-all" approach is helpful for reporting results in a relational database setting).

Generally, the results of applying a Gaussian Mixture Model to a transactional database results in a set of behaviors that are easily interpretable. The resulting clusters are understood as "segments" by marketing or merchandizing decision-makers. Each set of segment behaviors may be named by the user, and might form the basis for instance, of a promotional campaign. The model may also be maintained so that future transactions can be assigned a "score" according to the representative behavior involved. This allows the maintenance of databases for "intervention" analysis. An example of such a behavioral analysis might be: "Did the resulting promotional campaign increase the profitability of a given customer segment?"

CONCLUSION

This concludes the description of the preferred embodiment of the invention. The following paragraphs describe some alternative embodiments for accomplishing the same invention.

In one alternative embodiment, any type of computer could be used to implement the present invention. In addition, any database management system, decision support system, on-line analytic processing system, or other computer program that performs similar functions could be used with the present invention.

In summary, the present invention discloses a computer-implemented data mining system that analyzes data using Gaussian Mixture Models. The data is accessed from a database, and then an Expectation-Maximization (EM) algorithm is performed in the computer-implemented data mining system to create the Gaussian Mixture Model for the accessed data. The EM algorithm generates an output that describes clustering in the data by computing a mixture of probability distributions fitted to the accessed data.

The foregoing description of the preferred embodiment of the invention has been presented for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form disclosed. Many modifications and variations are possible in light of the above teaching. It is intended that the scope of the invention be limited not by this detailed description, but rather by the claims appended hereto.